Hans Journal of Computational Biology 计算生物学, 2015, 5(4), 49-57
Published Online December 2015 in Hans. http://www.hanspub.org/journal/hjcb
http://dx.doi.org/10.12677/hjcb.2015.54006

Hans汉斯

The Analysis and Visualization System of the Complementary and Matching Features of DNA Sequences

Wenjia Liu[1], Zhijie Zheng[2]

[1]School of Software, Yunnan University, Kunming Yunnan
[2]Key Laboratory of Information Security, School of Software, Yunnan University, Kunming Yunnan
Email: 8Avalon8@gmail.com

Received: Nov. 18th, 2015; accepted: Dec. 4th, 2015; published: Dec. 8th, 2015

Abstract

Due to the complementary symmetry properties of the double helix of DNA sequence and complicated space structure, exploring long-range DNA pieces matching characteristics, especially on the complementary relationship has important significance. In this paper, the substring-complementary string matching technique is used to predict the possible structure of hairpin structure and detect palindromic structure in DNA sequences. With using statistical measurement method to processing DNA, and comparing with the results of analysis, visualization of measured-date, large amounts of complex DNA sequences can be analyzed speedy. Through the results, the matching structure in selective DNA sequence does exist in long distance. Visualization methods, the analysis of the characteristics of the measuring model and extract visual mechanism given in the paper can provide a model and practice foundation for different DNA sequence data, and the application of visualization analysis of structure research.

Keywords

DNA Spatial Structure, Long Distance Matching, Substring Complementary-String Matching, Visualization

DNA序列互补匹配特征分析及可视化系统

刘文嘉[1]，郑智捷[2]

文章引用: 刘文嘉, 郑智捷. DNA 序列互补匹配特征分析及可视化系统[J]. 计算生物学, 2015, 5(4): 49-57.
http://dx.doi.org/10.12677/hjcb.2015.54006

[1]云南大学软件学院，云南 昆明
[2]云南大学软件学院信息安全重点实验室，云南 昆明
Email: 8Avalon8@gmail.com

收稿日期：2015年11月18日；录用日期：2015年12月4日；发布日期：2015年12月8日

摘 要

由于DNA序列双螺旋结构的互补对称性质以及空间上的复杂结构，探索长距离DNA片段的匹配特征尤其是在互补关系上有着重要的意义。本文利用子串－互补串匹配技术通过对DNA序列已有结构的检测分析着重对回文序列、发夹结构等可能存在的结构进行预测。通过统计测量的方法对DNA数据进行处理，对结果进行对比分析，将测量数据转换为可视图，对批量复杂DNA序列的提取特征数据进行可视化分析。通过结果图示，可以看到选择的DNA序列中的确存在着长距离匹配结构。文中给出的可视化方法，提出的分析测量模型以及提取的测量特征可视化机制能为后续不同DNA序列数据以及结构的可视化分析的应用研究提供坚实的模型和实践基础。

关键词

DNA空间结构，长距离匹配，子串－互补串匹配，可视化

1. 引言

自 1980 年的噬菌体 Φ-X174 实现完全测序，成为第一个测定的基因组以来，伴随着基因组学以及测序技术的发展，尤其是人类基因组计划从 20 世纪 90 年代启动到 2000 年人类基因组草图正式完成，再至 2005 年人类基因组计划的测序工作基本完成，已经有了难以计数的生物信息尚待人们挖掘与探索。面对海量的数据，在分子生物学的基础上，生物信息学[1]应运而生。生物信息学综合了各种理论和技术，主要对生物信息进行采集，处理，传播，分析和解释。而研究 DNA 序列的结构以及由 A、T、C、G 组成的碱基序列中的特征和规律是生物信息学中相当重要的课题。

在目前对于 DNA 序列的特征提取分析以及相似性比较的研究中，诸多有效的方法和工具正在不断涌现。如骆家伟等学者通过对 DNA 序列计算信息离散度来进行相似性分析[2]。同时利用可视化工具对 DNA 序列进行分析也是目前的一大热点。如白兰凤提出的利用二维图表示 DNA 序列，计算对应的距离矩阵并得到特征值以比较特征[3]。还有郑智捷等一批学者使用流密码的随机性检测方法以及变值逻辑体系对 DNA 序列进行可视化分析，并取得了一系列成果[4]-[7]。而随着目前计算机技术的发展与成熟诸如神经网络等技术也被应用在了如 DNA 序列分类中[8]。

而在本文中则区别于其他方法，将重点集中在 DNA 序列的对称和互补现象中，采用了子串－互补串匹配技术，针对特有的 DNA 序列结构如回文结构发夹结构，对 DNA 序列中的长距离匹配现象进行检索和分析，并通过可视化等一系列方式挖掘其中隐藏的信息，直接从序列一级结构层次上对可能出现的空间结构特征进行探索，为 DNA 序列精细特征的分析研究提供了一种新的方法。

2. 共轭对称及回文结构

在 DNA 序列中，碱基配对遵从着严格的互补对称关系，本文所研究的重点主要在 DNA 碱基序列上，为了找出具有特殊性的序列上的碱基片段，着重研究四种片段形式，分别是：片段本身，反向片段，互

补片段，反向互补片段。以片段 ACGTCA 为例，其四种形式如表 1 所示。

此四种形式的片段并非本文中第一次提出，同时也存在其它形式的片段，但在本文研究中，为了研究因互补配对造成的可能存在的空间结构，故而选取这种较基本的片段形式进行后续的搜索统计和观察。

同时，需要注意的是在诸多片段中，存在着一种较为特殊的片段，具体表现在该片段的反向互补片段是它的本身。如表 2 所示。

若出现表 2 情况的片段，则称其为片段与反向互补片段之间存在**共轭对称**关系。

而**回文结构**在生物基因组学中主要是指双链 DNA 中某段序列具有的反向重复的结构，当该序列的双链被打开后，可形成发夹结构。这段序列则被称为回文序列(Palindromic sequence)。

其特点是在该段的碱基序列的互补链之间正向与反向是相同的，例如

5’AAGCTT’3

3’TTCGAA’5

在双链中，可能如下方序列所示出现：

……C G A T T A C A G G C T A A G C T T T C C A G C G T A C A C G……

……G C T A A T G T C C G A T T C G A A A G G T C G C A T G T G C……

3. 系统和方法

本文建立的系统包括 3 个核心模块：特征统计模块、特征分析模块和可视化模块。处理过程为把若干段等长的序列分别进行特征统计，再从统计结果中选出具有代表性的特征，最后对选取得特征集合进行可视化，显示各序列中特征片段的分布特征。该系统的结果和处理过程如图 1 所示。

3.1. 数据处理模块

在诸多对 DNA 数据分析测量的研究中，使用了诸如数据挖掘，基于统计，基于概率等的各种方法来进行探索，而在对基因回文结构的研究探索中，期待能够发现回文结构并对其位置以及序列形式进行记录和统计分析，以得出此类结构是否存在某种特殊现象及分布规律，故而选择基于统计的方式来探索其

Table 1. Four kinds of corresponding relations of base fragment
表 1. 碱基片段的四种对应关系

片段本身	A C G T C A
反向片段	A C T G C A
互补片段	T G C A G T
反向互补片段	T G A C G T

Table 2. Conjugate symmetry of base pair
表 2. 碱基片段的共轭对称关系

片段本身	互补片段	反向互补片段
A A G C T T	T T C G A A	A A G C T T
C C T A G G	G G A T C C	C C T A G G
A C T A G T	T G A T C A	A C T A G T

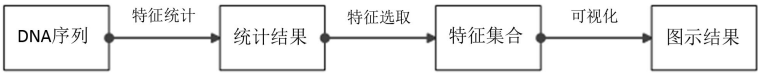

Figure 1. System model
图 1. 系统模型

特性。通过对对应序列之间的匹配，寻找两条序列之间是否存在由同种类型的碱基片段形成的回文结构，并观察其在同一数据集中的特征。最后用统一的数据结构保存得到的信息，为统计出现在数据集中的位置分布情况以及出现次数以及可视化提供数据。

本文中的匹配模式，即以数据集中的单条或成对 DNA 序列作为处理对象，在序列中通过滑动窗口的形式依次选择碱基片段，对于序列中选取碱基片段位置之后出现的对应形式碱基片段记录位置，从而实现匹配处理，得到关于数据集中每个出现碱基片段的所有形式的位置数据。匹配过程如图 2 所示。

模块的输入：

以 FASTA 格式保存记录的两组若干等数量等长度 DNA 序列集合，两组序列皆取自某基因左右两端各 500 bp 长 DNA 序列。每条编号的序列皆有另一条编号相同的序列与其对应。

模块的匹配处理：

从两组数据集中依次选取左右两条对应编号的 DNA 序列，使左右两条序列各自对自身序列进行匹配，记录片段位置信息，同时将左侧序列与右侧序列进行匹配，初始匹配片段长度从 3 开始，直至出现长度为 n 的碱基片段使得该序列中不存在有任意长度为 n + 1 的片段在对应序列中存在与其匹配形式的碱基片段。

以如下序列处理过程为例：

如步骤①所示，初始匹配片段长度为 3，选取前三个碱基 TAC 作为匹配片段，由前的匹配方法可知对应四中片段为 TAC，ATG，CAT，GTA。

通过往后检索，即可匹配到在右侧序列位置 5、位置 11、位置 20 出现了反向片段本身 CTT，未出现其它对应形式片段。

①

左：T A C T C G A C G AC T TCT A C G A C T A A G G C G T A T C C T C G A G

右：G C G TCATAAACATCTT G ACCATTGTGAATTCTAGAA C

记录片段 TAC 对应 4 种匹配形式的位置后，将匹配片段位置向下移一位，进行下一个片段的匹配，如步骤②所示，移至 ACT 片段处。

②

左：T A C T C G A C G AC T TCT A C G A C T A A G G C G T A T C C T C G A G

右：G C G TCATAAACATCTT G ACCATTGTGAATTCTAGAA C

如步骤③所示，当滑动窗口内选中的片段 ACT 已经在对应序列进行匹配处理过时，跳过该片段，滑动窗口向下移一位，移至 CTT 片段处。

③

左：T A C T C G A C G A C T T CT A C G A C T A A G G C G T A T C C T C G A G

右：G C G TCATAAACATCTT G ACCATTGTGAATTCTAGAA C

当长度为 3 的片段匹配到序列最后一位时，长度加 1，从序列前段开始再次匹配。如步骤 4 所示。

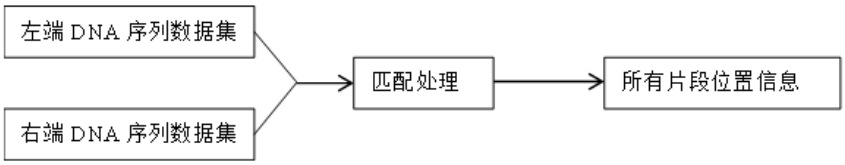

Figure 2. Matching model
图 2. 匹配模式模型

④

左：T A C T C G A C G AC T TCT A C G A C T A A G G C G T A T C C T C G A G

右：G C G TCATAAACATCTT G ACCATTGTGAATTCTAGAA C

当长度 n 等于序列长度的二分之一时，停止匹配。进行下一个序列的匹配处理。

直到数据集中的最后一对序列处理完毕后，该数据集的匹配处理过程结束。

模块的输出：

数据集所对应的结果数据集中，保存左侧序列自匹配结果，右侧序列自匹配结果以及左右对应异匹配结果所有有相关记录存在的序列的结果数据，每条序列的结果数据中通过匹配片段长度的不同将结果数据分离组织，不同长度匹配片段结果数据集中分别保存该长度下有匹配位置记录的片段结果数据。

处理匹配结果及分析：

数据集 1 的数据为 read 序列在 H2H 基因上 map 到两个位置，且位置相距 2 万 bp 以上的对称长度 500 bp 的左右序列结果如表 3 所示，数据集 2 的数据为随机在人类基因组上选取的相隔 2 万 bp 以上的对称长度 500bp 的左右序列，结果如表 4 所示。

在实际的处理过程当中，采用了多组数据进行分析处理，在通过对各组数据最终结果进行分析和比较后发现：

(1) 碱基片段“AAGCTT”在进行处理的数组 DNA 序列中出现频率较高，其反向互补是其本身，存在着共轭对称的情况，同时在生物学中是限制酶的切割位点。

(2) 对于所有数据集中的序列，每条序列中都存在长度至少为 8 或以上的片段有出现自身，互补，反向，反向互补四种形式之一的片段与其对应。同时可以发现存在对应片段情况的碱基片段平均长度为 11 到 13，可以推测如果出现回文结构发夹结构等现象互补长度不会太长。

Table 3. Processing result of dataset 1
表 3. 数据集 1 处理结果表

数据集中序列数量	3000 对
最长片段长度	165 bp
最短片段长度	8 bp
平均片段长度	12 bp
出现最多的片段	AAGCTT：1492 对序列中出现
	CCATGG：1184 对序列中出现
	ATTTTT：745 对序列中出现

Table 4. Processing result of dataset 2
表 4. 数据集 2 处理结果表

数据集中序列数量	3000 对
最长片段长度	76 bp
最短片段长度	8 bp
平均片段长度	12 bp
出现最多的片段	AAGCTT：1400 对序列中出现
	CCATGG：1036 对序列中出现
	TTTTTT：711 对序列中出现

(3) 在数据集 1 和数据集 2 中，可以发现得到的统计结果极其相似，而出现次数最多的 AAGCTT 以及 CCATGG 都是共轭对称同时它们都是某种限制酶。

3.2. 可视化分析方法

在获得片段位置相关数据并进行统计分析后，得到了每组数据中片段的出现情况，并重点观察分析了出现次数最多的片段，为了进一步对序列中有对应形式的片段位置的确认以及进行观察，需要建立一套能够更为直观地对其中的精细特征进行观察的可视化模型。模型的结构图如图 3 所示。

输入：匹配模块输出的位置结果及选定需可视化的片段；

处理过程：在输入数据集片段位置结果以及选定片段后，首先需要将所有选定长度的片段位置结果提取出来，再从中选定指定的片段对应形式的位置，将其可视化。

而在匹配过程中，以 X 轴表示序列长度，Y 轴表示不同序列，如数据集左右各 N 条序列，在第 n 组左右对应序列中，其长度为 L，片段 M 在自匹配左侧序列中位置 a，b (0 < a, b < L)存在对应形式，在异匹配右侧序列中位置 c，d (0 < c, d < L)存在有对应形式片段，则在坐标(n, −a)，(n, −b)，(n, c)，(n, d)作标志记录，当数据集中所有序列都进行过处理和显示后，生成图示，处理结束。

输出：针对自匹配所对应的片段位置图示以及自匹配和异匹配位置数据对应的片段对比位置图。

3.3. 可视化结果及分析

在获得片段位置相关数据并进行统计分析后，得到了每组数据中片段的出现情况，并重点观察分析了出现次数最多的片段，如图 4 至图 9 所示。

通过上述可视化图示，可以观察出下列现象及特征：

(1) 由数据集 1 中的图示可见，在数据集 1 中 ATTTTT 片段位置在序列中分布较为随机和均匀,而出现次数最多的 AAGCTT 以及其次的 CCATGG 在分布上具有十分明显的特征，它们在分布上自匹配和异

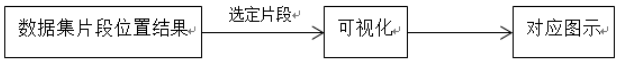

Figure 3. Visualization model
图 3. 可视化模型结构

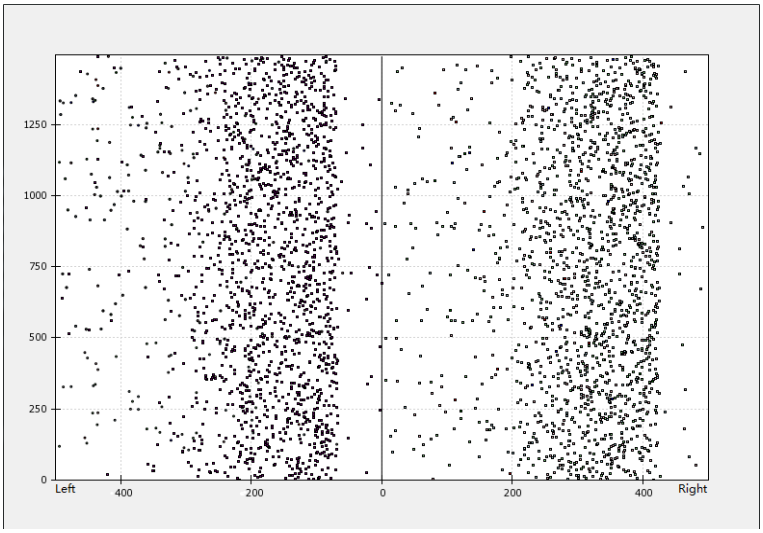

Figure 4. The matching results of AAGCTT in dataset 1
图 4. 数据集 1 中 AAGCTT 片段匹配结果图示

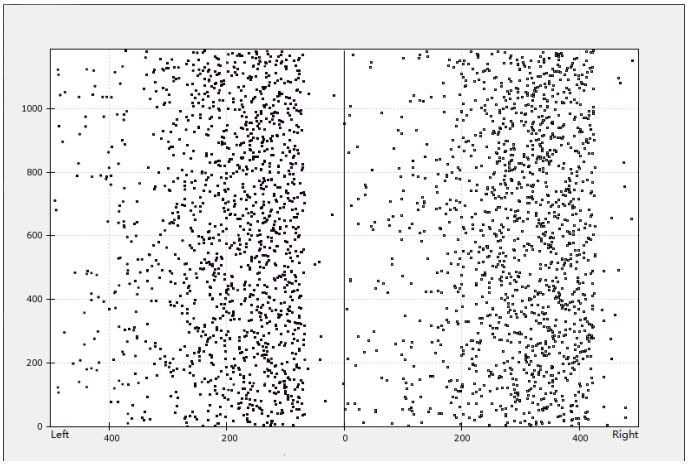

Figure 5. The matching results of CCATGG in dataset 1
图 5. 数据集 1 中 CCATGG 片段匹配结果图示

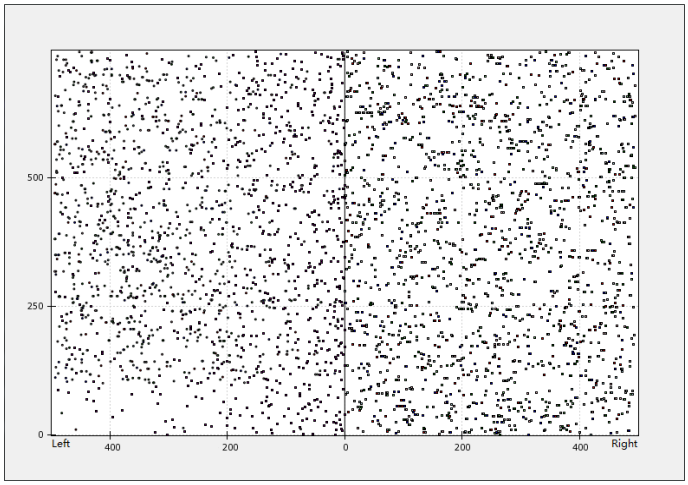

Figure 6. The matching results of ATTTTT in dataset 1
图 6. 数据集 1 中 ATTTTT 片段匹配结果图示

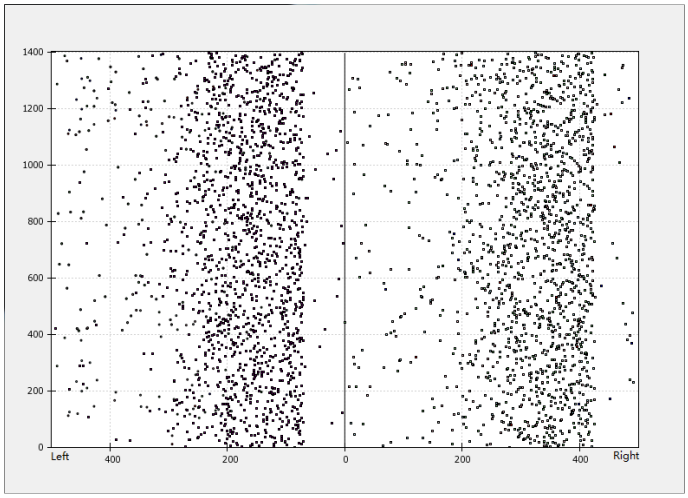

Figure 7. The matching results of AAGCTT in dataset 2
图 7. 数据集 2 中 AAGCTT 片段异匹配结果图示

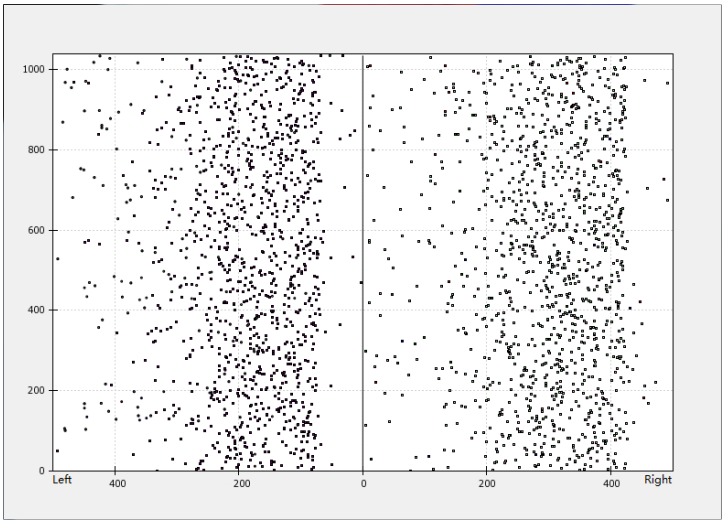

Figure 8. The matching results of CCATGG in dataset 2
图 8. 数据集 2 中 CCATGG 片段匹配结果图示

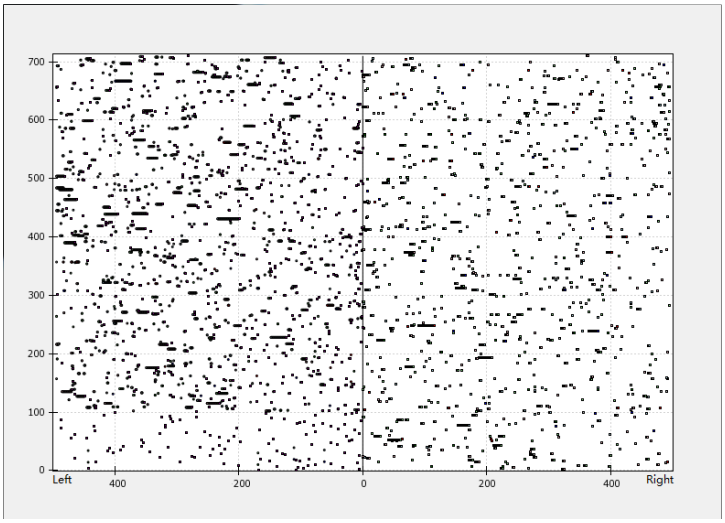

Figure 9. The matching results of TTTTTT in dataset 2
图 9. 数据集 2 中 TTTTTT 片段匹配结果图示

匹配都有很明显的分界线，在序列大概 50 bp 左右极少出现对应形式片段，大多出现在 50~250 bp 区间段，之后成彗星尾部状出现次数及频率逐渐减少。

(2) 从数据集 2 中的图示来看，呈现出了与数据集 1 中相似的特征与情况，AAGCTT 与 CCATGG 分布十分相似，而 TTTTTT 则无明显现象。

(3) 通过对数据集 1 和数据集 2 的对比中，两组数据呈现了相似的分布现象，出现次数最多的 AAGCTT 和 CCATGG 在另一组数据集中分布相似，说明了该特征不仅仅只局限于某一数据集。

(4) 在对选取的其它数据集如部分编码区与非编码区的分析可视化中，并未发现较明显的类似现象与特征，这说明数据集 1 和数据集 2 所在的 H2H 基因可能具有一定的特殊性。

4. 结论

本文通过对 DNA 精细特征的分析处理，对 DNA 序列中出现的四种片段的结构进行了分析和整理，

通过对专业处理后得到的数据集进行了匹配处理，得到了不同序列不同片段在数据集中的位置信息以及出现频率上的统计数据，在对这些结果进行分析后，确定了可视化的方法，选取了需要重点观察的片段对象，利用可视化方法对各数据集以及感兴趣的片段进行了自匹配可视化以及异匹配可视化展示。在最后的展示图示以及结果中，可以发现许多有意义的信息。从展示结果来看，部分片段如 AAGCTT 以及 CCATGG 在部分数据集中的分布以及数量可以为 DNA 序列的空间结构提供证据。同时更具体分布特性以及序列中的精细特征还可以进一步探索。

致　谢

感谢云南省教育厅重大专项(K1059178)、国家自然科学基金(61362014)的支持。

参考文献 (References)

[1] Cantor, C.R. and Lim, H.A. (1991) The First International Conference on Electrophoresis, Supercomputing, and the Human Genome: Proceedings of the April 10-13 Conference at Florida State University, Tallahassee, Florida. International Conference on Electrophoresis, Supercomputing, and the Human Genome, World Scientific.

[2] 骆嘉伟, 刘芳, 杨华. 基于信息离散度的 DNA 序列相似性分析[J]. 计算机应用, 2009, 29(1): 269-272.

[3] 白凤兰. DNA 序列的特征数值及相似性分析[J]. 数学的实践与认识, 2007, 37(18): 95-99.

[4] 张巍琼, 郑智捷. 基于不同产生机制的伪随机序列和 DNA 序列的随机性测量[J]. 成都信息工程学院学报, 2012(6): 548-555.

[5] 刘玉倩, 郑智捷. 编码和非编码 DNA 序列的可视化分析[J]. 计算生物学, 2014, 4(2): 20-31.

[6] 完竹, 郑智捷. DNA 序列一维分段测量分布可视化[J]. 云南大学学报(自然科学版), 2013(35): 1-6.

[7] 杜磊, 郑智捷. 在非线性函数下的 DNA 概率测量聚类分布[J]. 软件工程与应用, 2014, 3(3): 41-49.

[8] 敖丽敏, 罗存金. 基于神经网络集成的 DNA 序列分类方法研究[J]. 计算机仿真, 2012, 29(6): 171-175.